PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : C12Q 1/68, C12M 1/00

A1

(11) International Publication Number: **WO 96/36737**

(43) International Publication Date: 21 November 1996 (21.11.96)

(21) International Application Number: PCT/US96/07202

(22) International Filing Date: 16 May 1996 (16.05.96)

(30) Priority Data:
08/445,094 19 May 1995 (19.05.95) US

(71)(72) Applicant and Inventor: RABANI, Ely, Michael [US/US]; 4495 Vision Drive #1, San Diego, CA 92121-1942 (US).

(81) Designated States: AM, AT, AU, BB, BG, BR, BY, CA, CH, CN, CZ, DE, DK, EE, ES, FI, GB, GE, HU, IS, JP, KE, KG, KP, KR, KZ, LK, LR, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, TJ, TT, UA, UG, UZ, VN, ARIPO patent (KE, LS, MW, SD, SZ, UG), European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).

Published
With international search report.

(54) Title: PARALLEL MULTIPLEX POLYNUCLEOTIDE SEQUENCING

(57) Abstract

The present invention provides for the parallel detection of polynucleotide species size separated according to multiplex sequencing protocols, which are advantageously substituted for previous serial reprobing methods. This parallelism is realized by separately analyzing species handled by multiplex strategies in a length stratified manner. Stratification is accombplished by recovering size fractionated or size separated fractions, and examining these separately. The effort of analyzing many fractions is far offset by the advantages realized by parallel tag detection or probing; while a few hundred fractions may correspond to nested sets of fragments of up to tens or hundreds of bases in length, each fraction yields information about thousands to millions of contiguous sequences. Parallel tag detection is faborably accomplished by hybridization of multiplex sample with one or more stationary arrays of tag specific or sequence-segment specific probes. Information thus obtained is reassembled, according to the presence or absence of a tag, or base-dependent label associated with a tag, in each such fraction and the order of each such fraction in a contiguous series, generally by algorithmic means, to yield linear sequence information.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AM	Armenia	**GB**	United Kingdom	**MW**	Malawi
AT	Austria	**GE**	Georgia	**MX**	Mexico
AU	Australia	**GN**	Guinea	**NE**	Niger
BB	Barbados	**GR**	Greece	**NL**	Netherlands
BE	Belgium	**HU**	Hungary	**NO**	Norway
BF	Burkina Faso	**IE**	Ireland	**NZ**	New Zealand
BG	Bulgaria	**IT**	Italy	**PL**	Poland
BJ	Benin	**JP**	Japan	**PT**	Portugal
BR	Brazil	**KE**	Kenya	**RO**	Romania
BY	Belarus	**KG**	Kyrgystan	**RU**	Russian Federation
CA	Canada	**KP**	Democratic People's Republic of Korea	**SD**	Sudan
CF	Central African Republic			**SE**	Sweden
CG	Congo	**KR**	Republic of Korea	**SG**	Singapore
CH	Switzerland	**KZ**	Kazakhstan	**SI**	Slovenia
CI	Côte d'Ivoire	**LI**	Liechtenstein	**SK**	Slovakia
CM	Cameroon	**LK**	Sri Lanka	**SN**	Senegal
CN	China	**LR**	Liberia	**SZ**	Swaziland
CS	Czechoslovakia	**LT**	Lithuania	**TD**	Chad
CZ	Czech Republic	**LU**	Luxembourg	**TG**	Togo
DE	Germany	**LV**	Latvia	**TJ**	Tajikistan
DK	Denmark	**MC**	Monaco	**TT**	Trinidad and Tobago
EE	Estonia	**MD**	Republic of Moldova	**UA**	Ukraine
ES	Spain	**MG**	Madagascar	**UG**	Uganda
FI	Finland	**ML**	Mali	**US**	United States of America
FR	France	**MN**	Mongolia	**UZ**	Uzbekistan
GA	Gabon	**MR**	Mauritania	**VN**	Viet Nam

PARALLEL MULTIPLEX POLYNUCLEOTIDE SEQUENCING

Field of the Invention:

The invention relates to the field of molecular biology.

Related Art:

Due to the importance of the structure of genetic material to the understanding of biological mechanisms and genetic engineering, much effort has been focused on the development of techniques for the determination of the nucleotide base sequence of polynucleic acids, especially of DNA. At present there are only two well established basic methods for the determination of polynucleotide sequence. These are Maxam and Gilbert base specific chemical cleavage[1] and Sanger enzymatic chain termination[2]. Both of these methods serve to generate, from a sample of identical polynucleotide molecules, sub-populations of polynucleotide molecules with a particular one of the four naturally occurring nucleoside bases, according to treatment of the sample aliquot, at the 3′ terminal position. With both of these methods, each such differently terminated sample aliquot comprises a so-called "nested set" of populations of molecules of different length, up to the length of the input sample molecules, but identically having the specified 3′ terminal base composition. Thus, the length of each subpopulation within each such sample aliquot corresponds to a position of the base moiety, for which that sample aliquot has been treated, within the sample sequence. The nested sets produced by both of these methods are conventionally separated according to length by electrophoresis through gel matrices, generally of polyacrylamide composition. Generally, the sample is labeled with radiolabels or dye moieties such that appropriate detection techniques may localize the position of the bands which are produced by electrophoresis of different length molecules in the gel. By comparison of the pattern generated by such an electrophoresis step, the base sequence of the original sample is inferred. These methods are well known within the field of molecular biology.[3] To improve the rate at which sequence

information is accumulated, many different modifications of the original methods described above have been attempted or implemented. These include substituting different resolution media for polyacrylamide gels in the separation step[4], temporal or temporal and spatial separation of

fragments involving their departure from the separatory gel matrix (e.g. onto moving membranes or passed detectors), automation of fluid handling, reaction and electrophoresis steps with either conventional slab gels or gels confined to capillary tubes. Despite the improvements in efficiency attained by such modifications, established sequencing technologies are a rate limiting step in much molecular biological research and the several genome projects, including the Human Genome Project. In part, such limitations stem from the characteristics of electrophoretic separation, specifically the time involved and the decreasing resolution power for successively longer fragment length, which decreases approximately logarithmically. It is thus not generally feasible to sequence stretches of longer than 500 bases in length by gel electrophoresis, and much effort and care is required to extend this length beyond 1,000 bases at present.

Multiplex Sequencing:

One innovation which attempts to negotiate the limitations associated with gel electrophoresis is multiplex sequencing.[5] This method has been taught by G.M. Church in U.S. Patent Number 4,942,124 and further by G.M. Church and S. Kieffer-Higgins in U.S. Patent Number 5,149,625. These workers have noted that multiple samples may be separated within the same gel region (lane). When these are uniquely labeled with a tag sequence or otherwise uniquely detectable tag, the length-fractionated sequencing reaction products are then serially and uniquely detected according to tag identity. Thus, the pattern resulting from the electrophoretic fractionation of the sequencing products from each sample species is accessed by probing for the respective tag. Generally, this method is implemented by transferring or blotting the gel after electrophoresis onto a nitrocellulose or nylon membrane such that the size-separated polynucleotide molecules are deposited onto said membrane with preservation of spatial configuration of the resultant banding pattern. Usually a further

step, such as UV illumination of said membrane, is performed to further strengthen the immobilization of said polynucleotide molecules. Probing is generally performed with a radiolabelled, dye labeled or enzyme linked oligonucleotide complementary to the respective tag sequence. After image capture of the banding pattern thus revealed, the oligo probe is denatured from its cognate tag sequence and washed away from the membrane. The membrane is then serially reprobed in the same way until many or all distinctly tagged samples have been detected. Data from each probed sample is used to reconstruct sequence information in the same manner as with Maxam and Gilbert or Sanger sequencing methods. Beyond the reduction of electrophoresis time and effort per sample by this method, distinction with unique tags permits several samples to be pooled together, which enables several samples to be processed in only the four reactions specific to each base, significantly reducing handling steps. A drawback of this method has been the difficulty of automating the serial probing of membranes, though efforts have been made in this regard.[6,7,8]

Electrophoresis methods have been described in a volume entitled *Gel Electrophoresis of Nucleic Acids: A Practical Approach.*[9] These and other methods are taught by J. Sambrook, et al.[10]

Alternative Sequencing Methods:

Various novel methods for the determination of polynucleotide sequence have been proposed. These include substitution of mass spectroscopy[11] (including time-of-flight mass spectroscopy) separation for gel electrophoretic separation, exonucleolytic degradation with nucleotide transport and single molecule detection of the effluent[12], scanning probe microscopic visualization of polynucleotides at high resolution[13,14] with the aim of base discrimination within a single molecule, and various sequencing by hybridization methodologies. Of these, only the last appears at present to have significant practical potential, but significant obstacles and challenges remain.

Sequencing by Hybridization:

Two general variants of sequencing by hybridization methodology have been proposed. In the first, a sample under investigation is hybridized to an array of oligonucleotides situated on a surface[15],

where the composition of an array element is related to the position of said element within said array of oligonucleotides. For example, an array may comprise all possible 10-mer oligos, in which case there are over 10^6 array elements. An unknown sample is denatured and permitted to anneal with said array under stringent conditions. Said array is examined for the presence of bound sample molecules, generally by optical means. Ideally, the sequence composition of a sufficiently short sample molecule may be reconstructed by using information about which array elements have bound sample molecules and about the compositions of each of these array elements, with algorithms which overlap each of the short sequences (corresponding to the 10-mer oligos in this example) to produce a linear sequence. Two difficulties arise with this method. First, because DNA includes tandem and dispersed repeat elements (including transposons), a particular 10-mer, which is one of the more than 10^6 possible 10-mers may occur many times in naturally occurring sequences much shorter than a megabase.[16] Thus, this method will lead to branch points which ambiguate the sequence data thus obtained. Longer oligo probe libraries or arrays will contain more elements, which thus occur less frequently, but require more surface area and synthesis steps, and will further diminish the ability to discriminate between perfect and mismatched hybridization. Second, even for shorter oligos, such as 7-mers, hybridization is not perfect for all sequences under any given conditions, as is clear to those familiar with other probe hybridization techniques such as northern blot gel analysis; optimal conditions for a particular probe to form perfect hybrids and not form imperfect or mismatched hybrids vary in a sequence dependent manner, and it is not always simple to predict this optimum even where only a single probe is concerned. Thus, for any given array of all possible n-mers, only a subset of elements will display optimal binding discrimination characteristics under any set of particular conditions. Further, because of the formation of imperfect mismatches, discrimination between strongly and weekly binding probes must be considered rather than simply the success or failure of an array element to bind sample molecules.[17] Various attempts have been made to address these issues, with some success[18], but the technique is not yet competitive with established methods.

In the second variant of sequencing by hybridization methodology, the sample sequence (which is generally in the form of a phage or plasmid library) is immobilized, generally in the form of bacteriophage plaque or bacterial colony replicas, on a membrane, which is serially probed with a different labeled oligonucleotide of known sequence and length n. Here, because only one kind of hybrid is formed during each step, conditions may be closely optimized for homoduplex formation in each probing step. Plaques or colonies forming hybrids are noted. Repetition over the full combinatorial library of oligonucleotides of sufficient complexity relative to the complexity of the sample thus provides information about all of the n-mer sequences present in a particular plaque or colony. After each element in the full combinatorial library has been used to probe the sample sequence, linear sequence information may be reconstructed as in the first variant of sequencing by hybridization methodology. The two main drawbacks of this method are the requirement for synthesis of large probe libraries, which is much more laborious than the production of surface immobilized arrays by optically patterned deprotection[19], and that significant linear sequence information may only be reassembled only after many (thousands or more) probing steps.

The first variant of sequencing by hybridization is, however, proposed as a confirmatory sequencing method to identify any errors in sequence information obtained by other methods, and the same methodology is proposed as a means of classifying polynucleotides in unknown samples, for example, to detect the presence and type of polynucleotides present in clinical isolates. In this latter application, even where a particular polynucleotide displays a hybridization pattern which permits only ambiguous sequence reconstruction, a characteristic pattern which may serve as a highly specific signature may nonetheless be observed.

These methods have sought to eliminate electrophoresis steps or other separation steps to avoid the perceived limitations of these methods; thus, this line of work has effectively taught that electrophoresis is undesirable in novel sequencing approaches. Thus, there exist two contrasting approaches to improving sequencing methods: to avoid electrophoresis, or to specialize and refine electrophoresis

for improved performance according to established parameters and for use in the separation steps of established methodologies.

Object of the Invention:

It is an object of the present invention to gain the advantage of parallelism in the detection steps of multiplex sequencing and thus better exploit the parallelism gained in sample preparation and size fractionation steps gained with multiplex sequencing.

It is a further object of the present invention to provide rapid and inexpensive methods and means by which complex polynucleotide samples including large genomes may be analyzed with single base resolution.

It is a yet further object of the present invention to provide rapid and inexpensive methods and means for the high resolution analysis of many independent samples of genetic material, for example, from many individuals within a population of organisms.

Methodologies and means fulfilling one or more of these objectives would increase the rate of progress of much biological research and applications including recombinant DNA technology, and have relevance to both basic medical research and clinical medicine.

Summary of the Invention:

The present invention provides for the parallel detection of polynucleotide species size separated according to multiplex sequencing protocols, which methods are advantageously substituted for previous serial reprobing methods. This parallelism is realized by separately analyzing species handled by multiplex strategies in a length stratified manner. Stratification is accomplished by recovering size fractionated or size separated fractions, and examining these separately. The effort of analyzing many fractions is far offset by the advantages realized by parallel tag detection or probing; while a few hundred fractions may correspond to nested sets of fragments of up to tens or hundreds of bases in length, each fraction yields information about thousands to millions of contiguous sequences. Parallel tag detection is favorably accomplished by hybridization of multiplex sample with one or more stationary arrays of tag specific or sequence-segment specific probes. Information thus obtained is reassembled, according to the presence or absence or a tag, or base-

dependent label associated with a tag, in each such fraction and the order of each such fraction in a contiguous series, generally by algorithmic means, to yield linear sequence information.

Definitions of Terms:

Separatory coordinate: the degree of separation effected between two or more species by some fractionation or separation method. In conventional polyacrylamide gel electrophoretic sequencing, the separatory coordinate of each species is reflected by the distance that species has travelled in the gel matrix after some period of time under the action of an applied field; alternative separatory coordinates are time-of-flight in time-of-flight mas spectroscopic separations, or elution time in membrane-capture capillary gel electrophoresis. In all of these cases, the separatory coordinate at which a species is disposed after separation depends, for DNA molecules of length within an appropriate corresponding size range, upon molecular length. Here, separatory coordinate is directly related to separatory mobility, e.g. electrophoretic mobility in electrophoretic separations. This term and concept is general, and embraces, for example, elution of molecular species according to the composition of an applied elution buffer gradient applied in HPLC, wherein the composition of the applied elution buffer at the corresponding point along said gradient may be described as the separatory coordinate of molecules which are thence eluted.

Stratification: the division of a sample or sample population into groups or subpopulations on any basis esp. in order to select a sample from each group or in order to make separate observations on said groups or subpopulations; in the present disclosure, emphasis is placed on division of populations of molecules according to molecular size, length or mass.

Tag: a region or segment of a molecular species, the composition of which corresponds, preferably uniquely, to the identity of a said species in a population of molecules.

Probe: a molecule or portion thereof which exhibits binding to a tag, generally with predetermined specificity.

Multiplex: the processing or conveyance of multiple different entities (e.g. molecules or information,) generally simultaneously, through the same channel. For example, in electronics, two or more electrical or radiation signals may be combined or mixed onto the same channel or wire, transmitted by this channel or wire, and then "separated" by such means as frequency or amplitude filtering; temporal electronic multiplexing selects one of two or more signals to transmit and applies this one signal to a channel for transmission, while a receiving circuit may demultiplex this signal, if appropriate information is supplied, by switching it to one of two or more output channels. In DNA sequencing, this term initially denoted the performance of reactions on a mixture of several samples, which mixture is then separated in the same separatory channel (e.g. gel lane;) during later detection steps, information regarding clone identity is "demultiplexed" by serial reprobing (of a membrane to which size-separated species have been transferred in an ordered manner) with a single oligo- or poly-nucleotide probe, which selectively reveals a signal corresponding to the positions of molecules comprising the probed sequence. The term also refers to samples comprising mixtures.

Parallel: The simultaneous handling or processing of multiple different entities (e.g. molecules or information,) accomplished through the provision of multiple handling or processing means. Because multiple different handling or processing means are provided simultaneously, these are repeated and separated in space; thus, spatial parallelism (or repetition in space rather than time) is generally connoted by this term.

Serial: the repetition of the same operations (e.g. transmission, handling, processing or detection) in time on each element of a group comprising multiple different entities (e.g. molecules or information,) utilizing relatively few (and generally only one) channel, handling, processing or detection means at a time. Thus,

in serial reprobing of membranes prepared in multiplex sequencing protocols, each probed sequence separated in an individual gel lane is probed one after another in time.

It should be noted that there has been imprecision with which the terms "parallel" and "multiplex" have been used in recent years in the biological literature. Two examples will illustrate apparent interchanges of these meanings. S.P.A. Fodor and co-workers (Fodor, S.P.A.; et al.; **1993**. *Nature, 364*:555) have referred to the hybridization of complex samples of polynucleotides to oligonucleotide arrays using the term "multiplex." While such complex samples may be multiplexed samples in that they may comprise a mixture of different molecules (though for these workers, such different molecules must be of the same origin,) the essence of the advantage gained by the use of oligonucleotide or polypeptide arrays inheres in the fact that multiple hybridizations or binding reaction are performed *in parallel* on the same sample, such that detection may occur substantially *in parallel*. This is made clear when one considers that detection with these workers' methods requires that particular sample molecules be substantially separated in space before the detection step for each such molecular species to be correctly detected. Thus, the "multiplex" hybridizations or binding reactions performed by these workers actually comprises a plurality of individual hybridization or binding reaction separated in space. Most correctly, these assays *demultiplex* complex samples in a single parallel step to permit subsequent *parallel* detection steps. A second example is found in the work of E.S. Yeung and J.A. Taylor (Yeung, E.S.; Taylor, J.A.; **1994**. U.S. Patent Number 5,324,401) use the term multiplex to refer to the detection of polynucleotide fragments separated *in parallel* in an array of capillaries by capillary electrophoresis with *parallel* detection of signals from each capillary by a CCD camera. In spite of such incorrect or imprecise usages, it will be clear to one of ordinary skill in the art that there is a clear distinction between steps performed in parallel or signals detected in parallel and samples treated or separated in multiplex.

Description of the Invention:

The rate limiting step in multiplex sequencing technology is that of serial reprobing of size fractionated polynucleotides species. This rate is on the order of over one hour per membrane, which membrane will generally hold several tens of lanes and which membrane may be usefully reprobed generally fewer than one hundred times. Thus, each reprobing generally yields fewer than one hundred thousand bases of sequence data.

Size fractionated sequencing reaction products may be understood as a stratified sample. Rather than the usual method of spatial representation of this stratification within a gel or transferred from a gel to a membrane, complex populations of size fractionated polynucleotides may be stratified by size fractionation with separative isolation, such that each fraction may be independently analyzed. Each such fraction is collected in a manner such that it represents less than or equal to the resolved range (along a separation coordinate) of a unique length of polynucleotide molecules bearing a particular tag (i.e. such that sequencing reaction products corresponding to termination at two adjacent base positions are not present in any one fraction.) Fraction-wise recovery may be accomplished by dissection of a slab gel into narrow slices (containing one or less band) and electroelution from such slices, dissolution of similarly obtained slice to liberate entrapped polynucleotides, transfer to a membrane with selective and preferably ordered desorption of sample species from said membrane, or other methods of established art. A favorable separation comprises electrophoresis through a slab or capillary gel with collection of fractions exiting the border opposite, in the electric field, that at which the polynucleotide sample was applied to said gel. Liquid fractions thus collected may be directly applied to probe arrays.

Thus, rather than examining all length fractions bearing a particular tag sequence per serial probing step, a particular fraction is examined for the presence or absence of many tags, and optionally the identity of one or more labels which correspond to the identity of the terminal base moiety associated with each tag, in one massively parallel step. A continuous series of adjacent fractions is thus

examined with parallel detection such that stratified data regarding a length range of interest are collected (e.g. a few hundred bases of continuous sequence information per unique tag.) Sequence information is reconstructed by ordering the information obtained about tags in each fraction according to the linear order that represents the separation coordinate corresponding to the fractionation method used (e.g. a spatial coordinate for ordinary gel electrophoretic type resolution, a temporal coordinate for time-resolved or elution rate based electrophoretic separation, or time of flight in time-of-flight mass spectroscopy, etc.,.) The linear sequence that corresponds to a particular segment of sample polynucleotide material detected by means of a particular and unique tag is represented by the presence or absence of that tag in each size fraction (for "monochrome" or single base moiety specific sequencing reactions,) or by means of detection of the identity of a label associated, by communication within the molecular structure of the sequencing reaction product generally added during said sequencing reaction, with each particular unique tag, which label corresponds in a predetermined way to the identity of the terminal base of a sequencing reaction product (i.e. sample multiplexing both of different polynucleotide segments with associated tags and of sequencing reaction base specificity with associated terminal base specific labeling.)

Prior to sequencing reactions, for most embodiments of the present invention, input sample polynucleotides are preferably fragmented to manageable lengths (by standard methods such as shearing, mild digestion with appropriate endonuclease such as DNase I, or restriction enzyme digestion) to convenient length, denatured and permitted to hybridize to a probe array (or other solid, stationary or immobile phase immobilized probe population, such as a column, polymer matrix, etc.,) having a first probe population of identical composition to that to be used in the sequencing protocol. Unbound polynucleotides are collected with a wash step. These unbound nucleotides, by operational definition, do not contain sequences which hybridize to any element in the probe array. Thus, the unbound fraction is a polynucleotide sample depleted for fragments bearing sequences which might hybridize with any elements of a given probe population, which might complicate parallel detection or contribute to data loss. After a separate wash step to

remove any residual unbound polynucleotides, a second denaturation step is performed to recover the previously bound fraction. Said previously bound fraction may then be depleted against a second probe population (array or column) of sequence composition designed to be distinct from said first probe population, as was done against said first probe population. This fraction may then, similarly, be analyzed by the methods of the present invention using probe arrays consisting of said second probe population. Such depletion steps may be repeated, against different probe populations, as many times as necessary or desirable to maximize sequence information obtained for a sample.

Note that parallel detection of plural species (which may be termed multiplex detection) within a fraction may be performed serially on a fraction by fraction basis, or plural fractions may be examined simultaneously by provision of two or more suitable probe arrays. Each probe array may be examined for tag or sample binding by a single, dedicated detection means corresponding to each array, such as a CCD or video microscope, or plural probe arrays may be examined in sequentially by appropriate detection means. Detection rate may then be rate limiting with respect to data collection, but this will in any case be superior to serial reprobing of membranes. Further, a single array may be reused by the use of denaturation and wash steps, which may be verified by examination by the same detection means for residual label or bound polynucleotides. Detection methods are further discussed below. Considerations such as system cost, complexity and performance will primarily determine the parameters of such implementations, but it will be realized by those skilled in the relevant arts that both parallelism and pipelining may be exploited at every step by provision of appropriate apparata (e.g. plural sample and fluid handling channels, plural separation means, plural probe arrays, plural signal amplification and plural detection means,) in a manner measured to optimize system performance and minimize the idle time for any category of steps or apparatus components.

Note that a tag may either be a particular sequence (e.g. of a synthetic oligomer) added to the 5′ portion of a primer oligo used in Sanger enzymatic sequencing, which tag composition in some predetermined way corresponds uniquely to the sequence composition or as appropriate to the degenerate sequence composition, of said primer

oligo, or may be simply a short but unique sequence of the sample molecules themselves. This latter case poses a two-fold analogy to sequencing by hybridization in that detection occurs according to sample composition and that said detection of sample composition is exploited for parallel classification. Here, as concerns the polynucleotide sample, the "tag" is merely a portion of the sample itself, i.e. a classification. For purposes of this method and this invention, such a tag is nonetheless probed in an identical manner to tags which are added to a sample, and the tag may be defined merely as that which hybridizes to or associates with a particular probe. Probe populations are here either produced or chosen (including merely by selective observation , detection-recording, or sequence reconstruction of a group of probe specificities within a larger probe population or array) such that only probes that are unlikely to associate with more than one distinct sequence segment of the input sample are used. Probes which associate with more than one such sequence segment will yield data which correspond to the superimposition of the sequence data for all of those sequence segments to which said probes associate, and thus be readily identified, excluded from sequence data reconstruction, and where necessary, noted as information for use in further sample analysis.

Other tagging schemes are possible and do not depart from the essence of the present invention. For example, tagging of a primer may be accomplished through the chemical addition of one or more small molecule haptens or epitopes via linkers such that each class of primer (unique or degenerate) corresponds in a predetermined manner to a unique hapten or unique combination of haptens or epitopes. Detection is here accomplished with affinity arrays, for example, such as those described by S.P.A. Fodor et al.[20]

By such methods, and many obvious variations of these, parallel detection of tag identity tens or hundreds of thousands of different stretches of sequence information representing hundreds of contiguous bases may be obtained from a single gel lane or gel capillary.

Many of the various modifications of the original Maxam and Gilbert chemical sequencing methods and of the original Sanger enzymatic sequencing with chain terminator nucleotide analogues may be applied in combination with the methods of the present invention, as will be

obvious to those skilled in the arts of biochemistry, molecular biology, recombinant DNA technology, genetic engineering and related arts. Such variations include different priming schemes, variations in gel electrophoresis methods or conditions, substitution of alternative size-fractionation or size-separation methods, variations in labeling methods and schemes, variations in sequencing reactions, etc.,.

The present invention is generally relieved of most of the difficulties accompanying sequencing by hybridization because sequence data reconstruction does not directly depend on the sequence composition of probes or tags. When added to primers or sample molecules, tags are generally produced and/or chosen so as to not have sequence composition identical to any similar length sequence occurring in the sample. A set of tags of this kind and the corresponding probes used may thus be chosen so as to have hybridization specificity properties and discrimination versus heteroduplex binding which are optimal under a single set of conditions or a conveniently small group of conditions. Thus, a large array of probes may be produced so as to have optimal specificity under a manageably small number of conditions, i.e. such that parallelism is not compromised by diversity of hybridization stringency optima. Such issues have been discussed by W. Bains[21] Thus, such limitations of sequencing by hybridization methods are rather design considerations for the methods and means of the present invention.

Fractionation and Separation Methods and Means:

For use with the present invention, a fractionation and separation step or steps must provide single base resolution of sequencing reaction product lengths over a usefully long range (generally more than a few tens of bases and preferably more than a few hundred bases.) Such methods must further be compatible with probing of each such fraction in parallel for the species thus fractionated. Various chromatographic, electrophoretic and spectroscopic methods fulfill these criteria. A few are described, but it will be understood that many different methods are useful for purposes of the present invention

and may be employed without departing from the essence of the present invention.

Mass spectroscopy[22,23] is attractive because fractionation occurs over a short timescale compared to electrophoretic methods, and size separated products may conveniently be collected. Additionally, field induced or electrophoretic fractionation involving microfabricated[24] structures or devices, or other microstructured materials offers similar attractions of rapid fractionation because of the high degree of structural control, small distances required for transport with high length resolution, and high field strengths realizable. Such devices and materials shall be considered equivalents for purposes of the present invention.

Because of their widespread use within the associated fields, electrophoretic methods will be emphasized in this disclosure, but this choice is merely for illumination rather than limitation.

Size fractionation is generally a statistical process: transport of a species, i.e. a collection of identical molecules, depends upon many interactions of each molecule of said species with some other material or field. As such, there will be a characteristic mean and a characteristic distribution of mobility or of displacement over time characteristic of each species. Statistical mechanical processes dictate that each molecule has a very low but not necessarily negligible probability of experiencing transport or displacement of more than a few standard deviations from the mean and most of the distribution. This may be referred to as separatory scatter. Such scatter is one limitation on fractionation resolution; for purposes of the present invention, for a polynucleotide of length m to be within a useful range, there must be good separation between said polynucleotide of length m and those of lengths (m-1) and (m+1). Any given size fractionation or size separation method must separate species such that a sufficiently large useful range of polynucleotide lengths are obtained. Where a particular species consists of only a few molecules of identical length and sequence, such scatter may compromise detection. Therefore, a fundamental limit to the degree of multiplexing will be imposed by the fractionation method. A sufficiently large number of identical molecules must be provided to discriminate between erroneously scattered mobilities or displacements

and those falling within the bulk of the corresponding distribution, at or near the mean. Relative intensity or detection counts may be used to discriminate between occurrence of such infrequent events and events consistent with average behavior. In practice, only a few identical molecules will be necessary for sufficiently refined separation methods. In the case of electrophoretic methods, this may be simply estimated by the intensity ratio of a monodisperse band to the intensity of areas where no bands occur relative to an area of gel where no sample has been applied.

Sample Labeling and Detection:

Labeling may be accomplished, for example, with affinity labels and affinity association reaction with detection reagents, dye labels, fluorescent labels or radiolabels. Each such labeling method will have corresponding detection methods, obvious to those skilled in the relevant biochemical, molecular biological and instrumentation engineering arts. Different labeling means will be considered equivalent for purposes of the present invention, except where a specific means is explicitly identified in this disclosure.

Because a large number of different species may be separated and detected in parallel according to the present invention, it is preferred that methods for which a comparatively small number of molecules of each type and length are required for unambiguous detection be used. The reasons for this are twofold: at higher concentrations, some size fractionation methods suffer compromised resolving power, and the parallelism of detection entails that only a small portion of the input sample represent each species, which must still be detectable. Multiple methods meeting these criteria exist and may be considered equivalents for purposes of the present invention and disclosure. Only a few will be described.

Numerous methods capable of detecting a single molecule have been developed over recent years. These include affinity labeling and optical labeling.

In the field of biology, prominent among these are affinity labeling of particles such as gold colloids, or of fluorescently derivatized polymer beads. Samples are mixed or exposed to such beads, which are

retained only where sufficiently strong and long-lived affinity interaction occur. Such affinity interactions may be chosen to be highly specific, such that pronounced discrimination of target from non-target molecules is accomplished. Commonly used affinity interactions and labels include: streptavidin and biotin; digoxigenin and anti-digoxigenin antibodies; and, dinitrophenol and anti-dinitrophenol antibodies. A vast number of biological molecules also display such affinity interactions at various degrees of specificity. Nucleotide analogs, including analogs capable of terminating the template dependent enzymatic polymerization of polynucleotides, comprising affinity moieties derived from small molecule ligands or haptens have been prepared, are commercially available, and have been used in variations on Sanger enzymatic sequencing methods. Such analogs have been prepared from different nucleotide base moiety composition, including for sets of nucleotide analogs where each base moiety type is in communication with a different affinity ligand. Such analogs may be used with the present invention as follows: a single stranded DNA sample is subjected to Sanger enzymatic sequencing reactions with tagged primers in a single vessel with a set of four such distinctly affinity labeled chain terminator nucleotide analogs and all four of the ordinary deoxyribonucleotides at an appropriate ratio; the reaction products are denatured and subjected to capillary electrophoretic separation with fraction collection (i.e. separate aliquots are collected, as a function of electrophoresis time;) each aliquot is hybridized to an appropriate probe array specific for tag sequences only; each probe array is exposed to a stoichiometric excess (generally a slurry) of four distinctly fluorescently labeled polymer beads, each of which four types further bears one or more predetermined affinity group of a single type which binds specifically to exactly one of said distinctly affinity labeled chain terminator nucleotide analogs via the affinity label member; the presence, type and optionally also the intensity of fluorescence colocalized with each probe array element is recorded, generally by electronic means such as a color CCD device or video microscope, with data transfer to an electronic computer; computer algorithms are then used to reconstruct data thus obtained from the ensemble of fractions. In such an embodiment, the color of a bead localized to an array element indicates the identity of the 3'

terminal base moiety of polynucleotide molecules of a nucleotide length corresponding to the respective fraction analyzed with said array and comprising the tag sequence probed by said array element.

Example Protocol:

For example a genomic sample is fragmented by physical or enzymatic methods to yield a population of molecules with average fragment size of between about 600 and 1,000 base pairs. Such a fragmented genomic sample is then depleted of sequences homologous to a first tag-probe array to be used by hybridization under moderately stringent conditions. Non-binding material is collected, and then binding material is collected separately with a denaturing wash, for use in a similar procedure with a second tag-probe array comprising probes of composition different from those of said first tag-probe array. Non-binding material is then cloned by appropriate methods established in the art into a population of plasmid or phagemid vectors wherein each vector comprises a universal priming site immediately upstream of a variable tag sequence which is in turn upstream of a cloning insertion site. The set of said variable tag sequences is that set of sequences complementary to the set of probes of said first tag-probe array, to which said non-binding material did not bind. Such variable tag sequences (and the corresponding respective probe population) are chosen such that no tag sequence is likely to hybridize with any probe complementary to a different tag under moderately stringent conditions; the probe population comprises the set of probes complementary to all of said tag sequences. Examples of such libraries include those suitable for use with the original multiplex sequencing protocol of Church and Gilbert. The resulting sample is then transformed into a suitable host organism, the population of which is grown and optionally maintained. Template molecules prepared and/or purified in multiplex from such a library are split into four aliquots. Each of these four aliquots is primed by hybridization with a universal primer oligonucleotide derivated with one of four distinct labels. Each of said four aliquots is then subjected to a Sanger elongation reaction in the presence of one of four distinct chain terminating deoxynucleotide triphosphate molecules (e.g. to each aliquot is added one of: dideoxyadenine, dideoxycytosine, dideoxyguanine, dideoxythymine.)

Information regarding the correspondence between label type and chain terminator analog base moiety type (e.g. A, C, T or G) is recorded. Reaction products thus produced comprise a terminal base moiety the identity of which corresponds to the identity of said one of four distinct labels (according to the label derivated primer used with the respective chain terminator in the respective aliquot), with which they are in physical communication. Said four aliquots may then be combined into a single pool. Subsequent to Sanger reaction completion, a fluorescently labeled polynucleotide ladder of useful length range and increment (which preferably do not comprise any sequences homologous to any tag sequences used in such a process) may be added to said aliquots, or if they are combined to said pool. Said aliquots or said pool, comprising sequencing product molecules, are then size separated by separatory methods and means with single base separatory resolution over some useful size range (e.g. by electrophoresis through a denaturing polyacrylamide capillary gel or a capillary tube containing a solution of denaturant and high molecular weight polyethylene oxide [PEO,] separatory methods established within existing art, with successive collection of successive fractions as they exit said capillary under the action of an applied electric field.) Said fluorescently labeled polynucleotide ladder may facilitate determination of the optimal increment or delay between successive fractions, serving as an internal standard indicative of separation, and may be monitored by appropriate fluorescence detection means (as known within relevant arts;) information thus collected may be used to regulate and control, for example by electronic computer implemented algorithms in combination with a computer controlled fraction collector, the appropriate spacing or timing of collection of separate fractions. Fractions thus collected are then each applied separately and singly, under appropriate stringent hybridization conditions, to an oligonucleotide array each element of which comprises a number oligonucleotide probes of substantially identical sequence (probe array) complementary to preferably no more than one variable tag sequence; said oligonucleotide array preferably (though not necessarily) comprises the complete ensemble of oligodeoxyribonucleotide probes complementary to the complete ensemble of tags contained in said genomic library; the composition and identity

of each probe as a function of element position in said probe array is preferably known, but at least consistent for all arrays used in one procedure. Stringent washes may then be performed to remove unbound or poorly bound sample molecules. Depending on label type, steps developing or amplifying label signal are performed for all fractions or probe arrays (e.g. incubation of sample-hybridized probe arrays with affinity label receptor labeled latex beads, where the spectrum of such fluorescence for a given bead type corresponds distinguishably to the identity of said receptor, and where said receptor uniquely binds one of said four labels where said four labels are affinity labels.) Data regarding the association of label type or bead type with a probe-array element are then collected by optical means such as a CCD camera, and analyzed by computerized image analysis algorithms.

Once such data have been obtained for the ordered set of fractions (hence successive probe arrays,) sequence data for each clone corresponding to each tag and tag-probe are reconstructed: data obtained from successive probe arrays is sorted so as to arrange label identity data (and hence terminating base identity data) for each probe array element according to probe array element location within said probe array in the order by which successive probe arrays correspond to successive fractions and hence sample molecules differing in length by one base. Thus, data regarding a particular probe element will consist of information about the identity of labels in association with the corresponding probe element (of identical sequence composition) in each successive probe array, where such data are rearranged to the order corresponding to the order in which successive fractions applied to successive probe arrays were collected. For example, for rectangular 500x500 probe arrays comprising 250,000 unique probes, where 600 successive fractions (each denoted by an index F) differing in length by one base are collected and applied to 600 identical probe arrays, all labels associated with elements of the same index (referring to Cartesian coordinates, e.g. x-y coordinates, assigned to each element,) data collected from each array F for each element (x,y) may be placed in a 3 dimensional array with coordinates (x,y,F). Data are then reread for each probe array element (x_j,y_k) with F_m increasing from the minimum to maximum m for which useful data were obtained, as j and k are held constant. This data is the sequence of the clone

corresponding to the tag which is complementary to said particular probe (i.e. that located at (x_j,y_k)), and is of a read length $m_{max}-m_{min}+1$. Such a reconstruction process is performed for all probes. The maximum quantity of data which can be obtained is thus $(x_{max})(y_{max})(m_{max}-m_{min}+1)$, or 150,000,000 bases using the above parameters.

For example, the collected data:

(x_j,y_k)	F_m	Label Detected
(10,10)	36	A
(10,11)	36	C
(11,12)	36	G
(10,10)	37	A
(10,11)	37	G
(11,12)	37	A
(10,10)	38	G
(10,11)	38	C
(11,12)	38	T
(10,10)	39	G
(10,11)	39	C
(11,12)	39	C

correspond to the probed sequences:

Probe location	Base Position (F_m)
(x_j,y_k)	36 37 38 39
(10,10)	A--A--G--G
(10,11)	C--G--C--C
(11,12)	G--A--T--C

In other words, the sequencing reaction products binding to the probe located at coordinates (10,10) comprise the sequence 36-AAGG-39, those binding to the probe located at coordinates (10,11) comprise the sequence 36-CGCC-39, and (11,12) comprise the sequence 36-GATC-39. For fractionation by capillary gel exit, the length of collected species

increases with increasing time (and hence increasing fraction number) so that the sequences of the strands synthesized on input templates by Sanger reactions are 5'-36-AAGG-39, 5'-36-CGCC-39 and 5'-36-GATC-39, respectively.

Data thus reconstructed is further examined (for example by computer implemented algorithm) to detect cases where more than one clone has inadvertently bound to the same probe element; the identity of probes binding more than one clone may be noted.

Thus, the protocol of this example is most rapidly performed with a number of probe arrays equal to the number of separate length fractions collected and a similar number of label detection means.

In summary, the above example protocol comprises the steps:

1. Fragment a genomic sample;
2. Deplete said sample against the probe array which will be used to detect and discriminate tags;
3. Cloning into a population of vectors each comprising a universal priming site, a variable tag sequence and a cloning site;
4. Transform into a suitable host organism, grow as a library and purify template in multiplex;
5. Separate into four aliquots to each of which is added one of four distinctly labeled "universal" primers and a one of four distinct chain terminator analogs (one of: ddA, ddC, ddG and ddT);
6. Perform Sanger enzymatic reactions;
7. Optionally add a premade distinctly fluorescently labeled polynucleotide ladder;
8. Fractionate and collect ordered fractions F_m differing in length by one base;
9. Hybridize each fraction collected separately and singly to the probe array of step (2) above comprising probes complementary to the each variable tag of step (3) above;
10. As necessary develop or amplify label signals;
11. Detect identity of labels associated with each probe array element (x_j, y_k), and record these data for each probe array to

which each fraction F_m was applied, arranged into an array of data with coordinates (x,y,F);

12. Reconstruct clone sequence data by reading data for each probe array element (x_j,y_k) for increasing F_m (or $(x_j,y_k)_m$ for increasing m) with j and k held constant, and reorder accordingly;

13. Check data for each (x_j,y_k) for all F_m for ambiguity resulting from cross-hybridization.

Of course, not all probes in an array will yield useful probing and hence useful data for a particular sample, due for example to random duplications of tags within a sample. Rather, there will be some efficiency of probing, E. Thus arrays of n probes used to probe F fractions, with base specific labeling schemes, will yield (E)(n)(F) bases of sequence data. Alternatively, to obtain some quantity of data Q, arrays of probes size n=(Q)/(E)(F) must be used. E will generally be an empirical parameter, and will be affected by factors such as sample composition, probe ensemble composition, and binding or hybridization conditions. Generally, E>0.1 may be expected. Thus, provided steps are performed properly, to simultaneously detect and discriminate 100 different species in a single fraction, an array of 100/0.1=1,000 should generally prove adequate, to simultaneously detect and discriminate 10,000 different species in a single fraction, an array of 10,000/0.1=100,000 should generally prove adequate, and to simultaneously detect and discriminate 1,000,000 different species in a single fraction, an array of 1,000,000/0.1=10,000,000 should generally prove adequate. Improvements in effective probing efficiency will proportionately reduce these requirements.

Additional Embodiments:

Direct Sequencing of Genomic Material:

Genomic DNA may be digested with a particular restriction enzyme, optionally dephosphorylated, and ligated to a linker bearing an affinity or optical label, depleted against probe arrays, and subjected to appropriate sequencing reactions (e.g. either Maxam and Gilbert chemical sequencing methods, or subjected to partial exonuclease digestion followed by Sanger enzymatic sequencing methods.) Here, probing corresponds to hybridization to a small segment of the input genomic sequence on each fragment, where probes are selected or produced such that each restriction fragment is most likely to be probed by zero or one probe array element (i.e. one probe.) Such reactions are then separated and detected in multiplex and parallel, according to the general methods of the present invention. While this simple method is not likely to yield complete sequence information about a genome, it may yield a very large quantity of sequence information in a comparatively quite small number of steps. Thus it may be regarded as a broad spectrum data collection method.

Restriction Analysis, Sample Profiling and Disambiguation of Sequence Data:

As will be apparent to those skilled in the relevant arts, multiplex separation methods apply to other macromolecular samples, such as restriction digestion products of DNA samples. Multiplex detection, i.e. parallel detection on a fraction by fraction basis, is similarly applicable in accordance with the present invention. In this case, probing may detect a sequence occurring within a restriction fragment, and the fraction in which a probed sequence or (categorical) tag occurs indicates the length of the fragment thus identified.

Such a method may be used to rapidly profile a large DNA sample, and may further be employed to obtain data useful in the disambiguation of sequence data comprising branch points (divergences and convergences) which occur in sequencing methods which reassemble linear sequence information from information about shorter fragments, such as sequencing by hybridization. (Such branch points arise, for example,

in sequencing by hybridization, where a given n-mer occurs more than once in a given sample.[25] Internal tags comprising sequences found in stretches of sequence data near branch points are examined to detect length ranges between sites for the restriction enzyme used. When a tag sequence matches a sequence near an ambiguous branch, the distance in bases from said branch point to the restriction sites corresponding to the enzyme used are determined to obtain a possible length. When a possible length is consistent or similar to an observed restriction fragment (according to the restriction analysis method above, a branching may be selected as the correct sequence; alternatively, restriction analysis data obtained above may be used to eliminate a subset of branchings which are inconsistent with this restriction data.

For purposes of rapid sample profiling by restriction mapping, restriction digested samples may be fractionated into only a few fractions, possibly as few as only two fractions (e.g. one fraction containing all fragments 1 kilobase or shorter and another fraction containing all fragments longer than one kilobase,) which are then probed as above, e.g. for internal tags. In this minimal case of two fractions, data obtained will classify each probed sequence into one of two classes or categories. Most generally, where F fractions (corresponding to F different length ranges) are collected, and n probes are used for internal tag detection, F^n different sample profiles may be obtained from (F)(n) probes, without quantitating binding to each probe. If binding to probes is quantitated with Q levels distinguishable (which is possible with sufficiently sensitive detectors with sufficient dynamic range) this number of profiles is multiplied by a further factor of Q^n. Thus, the minimal case of two fractions can yield 2^n different detectable profiles without quantitation; for large n, this can be enormous discriminatory power which, because of the use of fractionation, is applicable to detecting length polymorphisms.

Devices Performing the Method of the Invention:

An automated device performing the method which is the essence of the present invention comprises a separatory channel, fraction collection means, probe binding means and detection means. Where fractions are not collected by direct transfer from said separatory

channel to said probe binding means, fraction transfer means must also be provided. Where samples are to be labeled according to procedures requiring multiple steps (e.g. affinity moiety labeling of samples followed in a later step by the binding of beads or particles comprising anti-ligands, receptors or other specific affinity groups which specifically recognize said affinity moiety,) means performing any subsequent of said multiple steps must likewise be provided. For example, this may comprise a first reservoir holding a slurry of affinity labeled beads or mixtures thereof, a pump, and an outflow which may be positioned to deposit a controlled quantity of affinity labeled beads onto the surface of said array, a second reservoir and a second pump for wash solutions to eliminate unbound beads from the surface of said arrays, and a third reservoir to collect waste solutions. Said first reservoir, which may contain a slurry, preferably is in communication with agitation means to prevent settling of particulates, or comprises integrated agitation means. Detection means will correspond in type to the label (e.g. fluorescent dye moiety, fluorescent bead, colloid,) and will favorably be combined with means which scan or otherwise move said detection means relative to said probe binding means, such that one detection means may examine many of the probe binding means for sample binding, for example such that one CCD may examine all probes used with particular fraction and further thus examine multiple fractions. All such means may be contained or enclosed within a single cabinet or case. While such an apparatus would simplify manual performance of the method of the present invention, it is preferred that such a device further comprise a computer system for controlling separation, sample fraction collection, sample fraction transfer, probe binding, detection, detector scanning relative to probe binding means, and to record collected data electronically. The same computer system or a different computer system may be employed to perform algorithms on collected data to reconstruct sequence data and align said sequence data.

A specific example comprises the following: one or more capillaries each comprising a sample input port and an outlet and filled with gel or low viscosity matrix separatory media, an electrical power supply to provide the electric field for electrophoresis, a plurality of oligonucleotide probe arrays situated on an electronically

controlled positioner (analogous to a fraction collector) which may translates each of said oligonucleotide probe arrays into position juxtaposed to the outlet of said capillary and may also translate said oligonucleotide probe arrays into position with provided means performing said subsequent labeling steps and into position for detection by a provided detection means, such as a color CCD camera; data collected by said CCD camera are transferred through a communications channel to a computer controller system provided to automate the device of the present example, and are stored with and/or without further algorithmic processing. Said oligonucleotide probe arrays preferably comprise on their surface, preferably integrated at one or more borders of said array, an electrode which serves to provide one pole of the electrophoretic field. Thus, fractions may be electrophoresed (more specifically, electroeluted) directly onto the target array, and the field thus established may sweep sample molecules across the surface of said array concurrent to fraction collection. Such a device preferably also contains one or more temperature control element for controlling binding reactions, e.g. hybridization stringency or dissociation temperature. Said oligonucleotide arrays are preferably translated such that they are contacted with said temperature control elements as sample fractions are applied or thereafter.

In general, detection means may be selected from a CCD camera, a cooled CCD camera, a photon counting CCD camera, a video microscope, a video confocal microscope, a scanning laser confocal microscope, though one of ordinary skill in the art will recognize that other detection means exist. Note that cooling a CCD camera improves signal-to-noise characteristics and thus permits either more rapid or more sensitive detection of optical labels. For such optical detection schemes, an appropriate light source must also be provided. Generally a laser or laser diode of appropriate wavelength may serve this function.

[1]Maxam, A.M. ; Gilbert, W.; **1977**. *Proc. Natl. Acad. Sci., U.S.A. 74*:560.

[2]Sanger, F.; Nicklen, S.; Coulson, A.R.; **1977**. *Proc. Natl. Acad. Sci., U.S.A. 74*:5463.

[3]See, for example, a standard laboratory manual: Sambrook, J.; Fritsch, E.F.; Maniatis, T.; **1989**. *Molecular Cloning: A Laboratory Manual*, Second Edition. Cold Spring Harbor Laboratory Press.

[4]For example, low viscosity media as taught U.S. Patent Number 5,374,527.

[5]Church, G.M.; Gilbert, W.; **1984**. *Proc. Natl. Acad. Sci., U.S.A. 81*:1991.

[6]Cherry, J.L.; *et al.*; **1994**. *Genomics, 20*:68.

[7]Church, G.M.; Kieffer-Higgins, S.; **1992**. U.S. Patent Number 5,149,625.

[8]Church, G.M.; **1992**. U.S. Patent Number 5,149,625.

[9]Rickwood, D.; Hames, B.D., eds.; **1982**. *Gel Electrophoresis of Nucleic Acids: A Practical Approach*. I.R.L. Press, Washington, D.C.,.

[10]Sambrook, J.; Fritsch, E.F.; Maniatis, T.; **1989**. *Molecular Cloning: A Laboratory Manual*, Second Edition. Cold Spring Harbor Laboratory Press..

[11]For example, see Beavis; R.C.; Chait; B.T.; **1994**. U.S. Patent Number 5,288,644.

[12]Jett, J.H.; Keller, R.A.; *et al.*; **1990**. U.S. Patent Number 4,962,037.

[13]For examples, see: Driscol, R.J.; Youngquist, M.G.; Baldeschwieler, J.D.; **1990**. *Nature, 346*:294;

Dunlap, D.D.; Bustamante, C.; **1989**. *Nature, 342*:204;

Allison, D.P.; *et al.*; **1992**. *Proc. Natl. Acad. Sci., U.S.A. 89:*:10129;

and for a review, see: Stine, W.B.; Rabani, E.M.; Smith, D.W.; *et al.*; **1995**. "Imaging of DNA Bases and Oligodeoxynucleotides using Scanning Tunneling Microscopy," in: *Frontiers in Biological Physics*. Academic Press, San Diego.

[14]Lindsay, S.M.; Philipp, M.; **1992**. U.S. Patent Number 5,106,729.

[15]Fodor, S.P.A.; *et al.*; **1991**. *Science, 251*:767.

[16]Bains, W.; **1994**. *Genome Analysis, Techniques and Applications, 11(3)*:49.

[17]Broude, N.E.; Smith, C.L.; Cantor, C.R.; *et al.*; **1994**. *Proc. Natl. Acad. Sci., U.S.A. 91:*:3072.

[18]Broude, N.E.; Smith, C.L.; Cantor, C.R.; *et al.*; **1994**. *Proc. Natl. Acad. Sci., U.S.A. 91:*:3072.

[19]Fodor, S.P.A.; *et al.*; **1991**. *Science, 251*:767.

[20]Fodor, S.P.A.; *et al.*; **1991**. *Science, 251*:767.

[21]Bains, W.; **1994**. *Genome Analysis, Techniques and Applications, 11(3)*:49.

[22]For a review of biological applications of mass spectrometry, see Siuzdak, G.; **1994**. *Proc. Natl. Acad. Sci., U.S.A.* *91*:11290.

[23]Beavis; R.C.; Chait; B.T.; **1994**. U.S. Patent Number 5,288,644.

[24]Woolley, A.T.; Mathies, R.A.; **1994**. *Proc. Natl. Acad. Sci. U.S.A.*, *91*:11348.

[25]Bains, W.; **1994**. *Genome Analysis, Techniques and Applications, 11(3)*:49.

whereforth, I claim:

Claims:

1. Method for both the detection and discrimination of two or more distinct molecular species of similar separatory mobility but different composition in a single step, subsequent to a separation or fractionation step in which the collectivity of said two or more molecular species of similar separatory mobility were separated from other molecules according to differences in mobility in the same separatory channel or distinct portion thereof.

2. Method for the determination of the base sequence composition of a polynucleotide sample according to claim 1 where said two or more molecular species are polynucleotide molecules produced in chemical or enzymatic sequencing reactions, and said two or more molecular species are substantially separated in said separation or fractionation step from other polynucleotide molecules of different length or separatory mobility also produced in said chemical or enzymatic sequencing reactions.

3. Method for the reconstruction of data concerning the composition of polynucleotide samples according to ordered data obtained by the method and means of claim 2.

4. The method of claim 2 where said size separation step or steps comprise one or more of: gel electrophoresis; capillary gel electrophoresis; gel electrophoresis with transfer to a membrane; gel electrophoresis with transfer to a moving membrane followed by selective desorption of sample molecules from said membrane; mass spectroscopy; electrophoresis through low viscosity media; and separation means comprising microfabricated or microstructural devices.

5. A method for the depletion of a polynucleotide sample against a stationary probe population comprising the steps of contacting said polynucleotide sample with said array, permitting

hybridization to occur to completion, and collecting the unbound portion of said polynucleotide sample.

6. A method according to claim 4 further comprising the additional step of recovering the population of molecules retained by said stationary probe population.

7. The method and means of claim 1 comprising diverse probe arrays situated on a substantially solid surface or stationary phase.

8. The method and means of claim 7 where detection and discrimination comprises a step binding to affinity labeled particles, colloids or polymer beads to affinity labeled sample molecules and where said sample molecules are subsequently or were previously bound to a probe of substantially known composition or identity.

9. A method according to claim 1 where said two or more distinct molecular species are restriction fragments of DNA.

10. A method for disambiguating branches in reconstructed sequence data according to claim 9 where internal tags comprising sequences found in streteches of sequence data near branch points are examined to detect length ranges between sites for the restriction enzyme used.

11. Method and means for the determination of polynucleotide sequence composition according to claim 2 where said polynucleotide molecules further comprise distinct variable tag sequences.

12. Method and means for the determination of polynucleotide sequence composition according to claim 8 where said affinity labels correspond in identity to the identity of the terminal base of each sample molecule.

13. The method of claim 1 where said molecular species are labeled with one or more optical label moieties.

14. The method of claim 1 where two or more separatory channels are used to separate two or more samples into two or more ordered ensembles of fractions.

15. The method of claim 2 where detection and discrimination of 100 or more distinct molecular species of substantially similar separatory mobility but generally different composition occurs in said single step of said detection and discrimination.

16. The method of claim 2 where detection and discrimination of 10,000 or more distinct molecular species of substantially similar separatory mobility but generally different composition occurs in said single step of said detection and discrimination.

17. The method of claim 2 where detection and discrimination of 1,000,000 or more distinct molecular species of substantially similar separatory mobility but generally different composition occurs in said single step of said detection and discrimination.

18. Method for labeling of two or more distinct molecular species for simultaneous detection and discrimination of said two or more distinct molecular species bound to a diverse probe array in a detection step, comprising a prior labeling step in which affinity labeled beads, fluorescent polymeric affinity labeled beads or affinity labeled colloids are contacted with the sample volume comprising said two or more distinct molecular species, and permitted to bind to affinity moieties comprised within the structure of said two or more distinct molecular species.

19. The method of claim 18 where two or more distinct affinities are each provided respectively on the surfaces of one of two or more types of particles which may be polymeric beads, fluorescently labeled polymeric beads, or colloids, each of said two or more types of particles being perceptibly different in color or size or color and size from all other said two or more types particles such that there exists a correspondence between the identity of

the one of said distinct affinities on a particle and the detectable appearance of the one of said two or more types of particles on which it is situated, and said two or more distinct affinities are chosen to specifically bind and hence discriminate two or more affinity labels comprised within the structure of said two or more distinct molecular species.

20. A method according to claim 1 comprising the steps of:

(a) fractionating a complex molecular sample according to separatory mobility in a single separatory channel;

(b) collecting two or more distinct fractions;

(c) probing each of said two or more distinct fractions obtained in step (b) with two or more distinctly identifiable probes each capable of specifically recognizing a distinct tag, where said two or more distinctly identifiable probes are contacted with one of said distinct fractions simultaneously, and as necessary washing away unbound molecules;

(d) detecting two or more probe-sample binding complexes obtained in step (c) simultaneously; and,

(e) recording information obtained in step (d), preferably electronically, according to probe identity and fraction separatory coordinate.

21. A method according to claim 20 where said molecular sample comprises polydeoxyribonucleotide molecules.

22. A method according to claim 20 where said molecular sample comprises polyribonucleotide molecules.

23. A method according to claim 22 where ten or more fractions are collected through steps (a) and (b).

24. A method according to claim 23 where said complex molecular sample comprises polynucleotide sequencing reaction products.

25. A method according to claim **24** where said polynucleotide sequencing reaction products are labeled.

26. A method according to claim **24** where said polynucleotide sequencing reaction products are labeled according to terminal base moiety identity.

27. A method according to claim **26** where said polynucleotide sequencing reaction products are labeled with labeling moieties selected from the group consisting of: dye moieties, fluorescent dye moieties, affinity moieties.

28. A method according to claim **27** further comprising a step of binding labeling particles to said affinity moieties comprised within the structure of two or more said polynucleotide sequencing reaction products where distinct affinities, receptors or anti-ligands are each provided respectively on the surfaces of one of two or more types of particles which may be polymeric beads, fluorescently labeled polymeric beads, or colloids, each of said two or more types of particles being perceptibly different in color or size or color and size from all other said two or more types particles such that there exists a correspondence between the identity of the one of said distinct affinities on a particle and the detectable appearance of the one of said two or more types of particles on which it is situated, and said two or more distinct affinities are chosen to specifically bind and hence discriminate two or more affinity moiety labels comprised within the structure of said two or more said polynucleotide sequencing reaction products.

29. A method for determining the sequence composition of plural polynucleotide molecules according to claim **24** and further comprising the step of reconstructing ordered sequence data according to the detection of sample molecules or the detection of base-specific labels in association with particular individual probes ordered according to the separatory coordinate in which said association occurs.

30. A method according to claim **20** where said fractionating step is performed with one or more separatory means selected from the group consisting of: a slab gel; a capillary gel; a low viscosity media in a capillary tube; a mass spectroscopic channel; a microfabricated channel; a microfabricated channel containing gel media; a microfabricated channel containing low viscosity separatory media.

31. An automated device for analyzing complex molecular samples comprising a separatory channel, an electrical power supply, a positioner or fraction collection means, two or more probe arrays, and a detection device.

32. An automated device according to claim **31** further comprising a computer control system.

33. An automated device according to claim **31** further comprising one or more reservoirs for labeling and wash reagents.

34. An automated device according to claim **31** further comprising one or more temperature control element for controlling binding reactions.

35. An automated device according to claim **31** where said detection device comprises an optical device selected from the group consisting of: a CCD camera, a cooled CCD camera, a photon counting CCD camera, a video microscope, a video confocal microscope, a scanning laser confocal microscope.

36. An automated device according to claim **31** substantially contained or enclosed within a single cabinet or case.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US96/07202

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) :C12Q 1/68; C12M 1/00
US CL :435/6, 287
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435/6, 287

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

APS, CAPLUS, MEDLINE, EMBASE, BIOSIS, INPADOC

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X --- Y	SANGER et al. A Rapid Method for Determining Sequences in DNA by Primed Synthesis with DNA Polymerases. J. Mol. Biol. 1975, Vol. 94, pages 441-448. See pages 443-444, plates I and II.	1-4, 7, 9, 10, 13, 14, 15 --- 16, 17
X --- Y	US 4,942,124 A (G. M. CHURCH) 17 July 1990 (17.07.90). See figures 4, 5 and 7, columns 1-3 and 6-10.	1-4, 7-15, 18-21, 23-30 --- 16, 17, 22
X --- Y	US 5,409,811 A (TABOR ET AL.) 25 April 1995 (25.04.95). See columns 1-7 and figure 1.	1-4, 6, 7, 9-11, 13-15, 31-36 --- 16, 17

[X] Further documents are listed in the continuation of Box C. [ ] See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E"	earlier document published on or after the international filing date	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&"	document member of the same patent family
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search
06 AUGUST 1996

Date of mailing of the international search report
27 AUG 1996

Name and mailing address of the ISA/US
Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231
Facsimile No. (703) 305-3230

Authorized officer
AMY ATZEL
Telephone No. (703) 308-0196

Form PCT/ISA/210 (second sheet)(July 1992)*

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US96/07202

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 5,237,016 A (GHOSH ET AL.) 17 August 1993 (17.08.93). See columns 1-5.	5, 6